Pneumonia Detection from Chest X-Ray Images using Machine Learning

Machine Learning
Final Report

Jatesh Joshi
12/17/2025

Table Of Content

4.1 Introduction

Pneumonia is a serious respiratory infection that affects millions of people worldwide and can be life threatening if not diagnosed and treated early. Chest X-ray imaging is one of the most common diagnostic tools used by clinicians to identify pneumonia related symptoms and problems in the lungs. However, manual interpretation of X-ray images can be time consuming and also runs the risk of human error, especially in busier clinics or hospitals that might be understaffed as well.

Recent advances in machine learning have enabled automated image analysis techniques that can assist medical professionals by improving diagnostic accuracy and efficiency. In this project, machine learning models are applied to classify chest X-ray images into two categories: normal and pneumonia-related lung opacity. The goal is not only to achieve strong predictive performance, but also to analyze how different modeling approaches behave when applied to medical image data.

This project explores both deep learning and traditional machine learning techniques, allowing for a comparison between models that learn features directly from images and models that rely on engineered feature representations. The results are evaluated using multiple performance metrics, with a focus on understanding trade-offs such as sensitivity, specificity, and class imbalance.

4.2 Dataset Description

4.2.1 Dataset Source

The dataset used in this project is an adapted version of the *Chest X-Ray Images (Pneumonia)* dataset originally published on Kaggle by Mr. Paul Mooney. The adapted dataset redistributes observations across training, validation, and testing sets to provide a more balanced experimental setup while preserving the original class structure.

The dataset consists of labeled chest X-ray images categorized into two classes: *Normal* and *Pneumonia (lung clarity/opacity)*. All images are stored in separate directories corresponding to their dataset split and class label, making the dataset organized and suitable for supervised learning tasks.

The dataset is publicly available on Kaggle at:
https://www.kaggle.com/datasets/pcbreviglieri/pneumonia-xray-images

4.2.2 Class Distribution and Dataset Splits

Dataset Split	Normal	Pneumonia (Opacity)	Total
Training	1082	3110	4192
Validation	267	773	1040
Test	234	390	624
Total	-	-	5856

Table 1: Distribution of chest X-ray images across training, validation, and test sets.

The dataset was split into training, validation, and test sets prior to model development to ensure a fair evaluation of model performance. The class distribution shows a clear imbalance, with pneumonia cases significantly outnumbering normal cases across all splits. This imbalance reflects real-world clinical data, where abnormal cases are more prevalent in diagnostic datasets. Rather than artificially balancing the data, the imbalance was intentionally
preserved to analyze how different models respond to realistic class distributions.

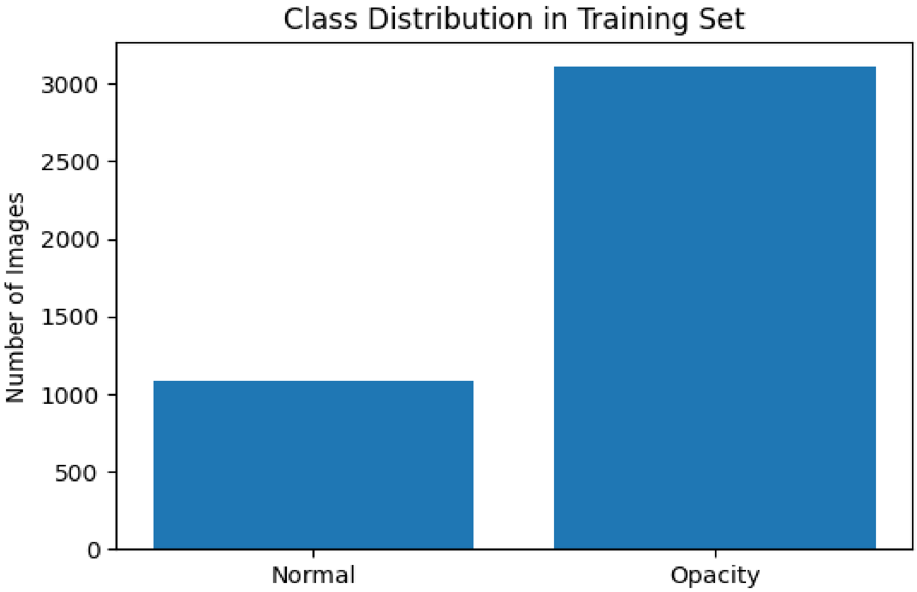


Figure 1: Class distribution of normal and pneumonia cases in the dataset.

4.3 Dataset Preprocessing and Design Choices

4.3.1 Dataset Cleaning

The dataset used in this project was already well curated and organized, with images stored in class-labeled directories and no missing or corrupted files. As a result, no heavy data cleaning procedures such as missing-value handling or label correction were required since most of the data was x-ray images. This allowed the focus of the project to remain on model selection, evaluation, and interpretation rather than extensive preprocessing.

4.3.2 Image Preprocessing

Prior to model training, all images were resized to a fixed resolution and normalized by scaling pixel values to the range [0, 1]. Resizing ensured consistent input dimensions across all models, while normalization improved numerical stability during training.

For deep learning models, images were processed in their spatial form, whereas for traditional machine learning models, images were flattened into one-dimensional feature vectors to enable the use of standard classifiers and all of this was done in one markdown file in google colab.

4.3.3 Handling Class Imbalance

The dataset exhibits a noticeable class imbalance, with pneumonia cases occurring more frequently than normal cases. Rather than applying resampling techniques or class rebalancing, the imbalance was intentionally preserved. This decision was made to reflect realistic clinical data distributions and to analyze how different models perform under such conditions. In real-world medical settings, chest x-rays are often ordered after patients begin exhibiting symptoms, which naturally increases the likelihood of positive findings. Model evaluation therefore focused on metrics beyond overall accuracy, including precision, recall, and class-specific performance measures.

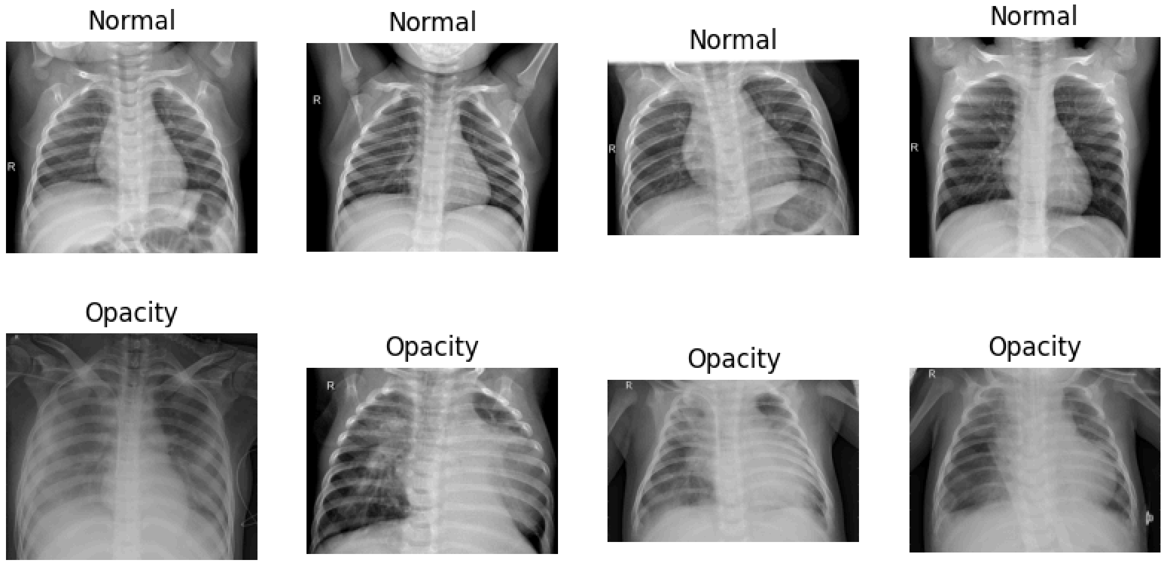


Figure 2: Example chest X-ray images from the dataset illustrating the visual differences between normal lungs and lungs exhibiting opacity associated with pneumonia.

This figure highlights the visual complexity of the classification task. While opacity patterns may be apparent in some cases, variations in lighting, contrast, and anatomical structure make automated classification challenging, motivating the use of machine learning approaches in this case.

4.4 Models and Experimental Setup

This project explores three supervised learning models of increasing complexity to classify chest X-ray images as normal or pneumonia-related opacity. The models were selected to allow comparison between deep learning and traditional machine learning approaches, as well as to analyze how different modeling strategies handle high-dimensional image data and class imbalance. All models were trained using the same dataset splits to ensure fair and consistent evaluation.

4.4.1 Model 1: Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) was implemented as the primary deep learning model due to its effectiveness in image classification tasks. CNNs are designed to automatically learn spatial features from images, making them well-suited for medical imaging applications.

The network architecture consisted of multiple convolutional layers followed by pooling layers to extract hierarchical features, and fully connected layers for final classification. A binary cross-entropy loss function was used, along with an adaptive optimization algorithm. Images were processed in batches, and a validation set was used to monitor performance during training.

Hyperparameter choice:

Initial training was performed using a fixed architecture and a predefined number of training epochs. To reduce overfitting and improve generalization, a limited hyperparameter adjustment was applied by modifying the number of training epochs based on validation performance.

4.4.2 Model 2: PCA Based Traditional Classifier

In addition to deep learning, a traditional machine learning model was implemented as a baseline comparison. Since classical classifiers cannot directly process high-dimensional image data, chest X-ray images were first flattened into one-dimensional feature vectors. Feature scaling was applied to standardize the data, followed by Principal Component Analysis (PCA) to reduce dimensionality while retaining 95% of the original variance.

This approach allowed the model to operate on a compact representation of the image data while preserving the most informative components. The PCA-based model serves as a baseline to evaluate the benefits of spatial feature learning provided by CNNs.

Hyperparameter choice

The Logistic Regression model was trained with L2 regularization to reduce overfitting in the high-dimensional feature space. The regularization strength was selected empirically to balance model simplicity and classification performance. No extensive hyperparameter tuning was performed, as the purpose of this model was to serve as a baseline for comparison.

4.4.3 Model 3: Random Forest

An ensemble learning approach was explored using a Random Forest classifier trained on the PCA-transformed features. Random Forests combine multiple decision trees to capture non-linear relationships in the data and often provide improved robustness compared to single classifiers. This model was included to assess whether ensemble methods could improve classification performance over simpler traditional models while remaining computationally efficient.

Hyperparameter choice

> The Random Forest model was trained using a fixed number of decision trees. The number of estimators and maximum tree depth were selected empirically to balance model performance and computational cost. Extensive hyperparameter tuning was not performed, as the objective was to evaluate ensemble behavior rather than optimize performance.

4.5 Evaluation Metrics

To evaluate model performance, multiple metrics were used to capture different aspects of classification quality. Since the dataset is imbalanced, relying solely on accuracy could lead to misleading conclusions. Therefore, a combination of threshold-based and probability-based metrics was selected to provide a comprehensive assessment of each model's behavior.

Accuracy: Measures the proportion of correctly classified images out of the total number of samples.

Precision (Positive Predictive Value, PPV): Represents the proportion of predicted pneumonia cases that are truly pneumonia, reflecting how reliable positive predictions are.

Recall (True Positive Rate, TPR): Measures the proportion of actual pneumonia cases that are correctly identified by the model, which is particularly important in medical diagnosis tasks where missed detections can have serious consequences.

F1-score: The harmonic mean of precision and recall, providing a balanced metric when class distributions are uneven.

True Negative Rate (TNR): Measures the proportion of normal cases that are correctly classified as normal.

Negative Predictive Value (NPV): Represents the proportion of predicted normal cases that are truly normal.

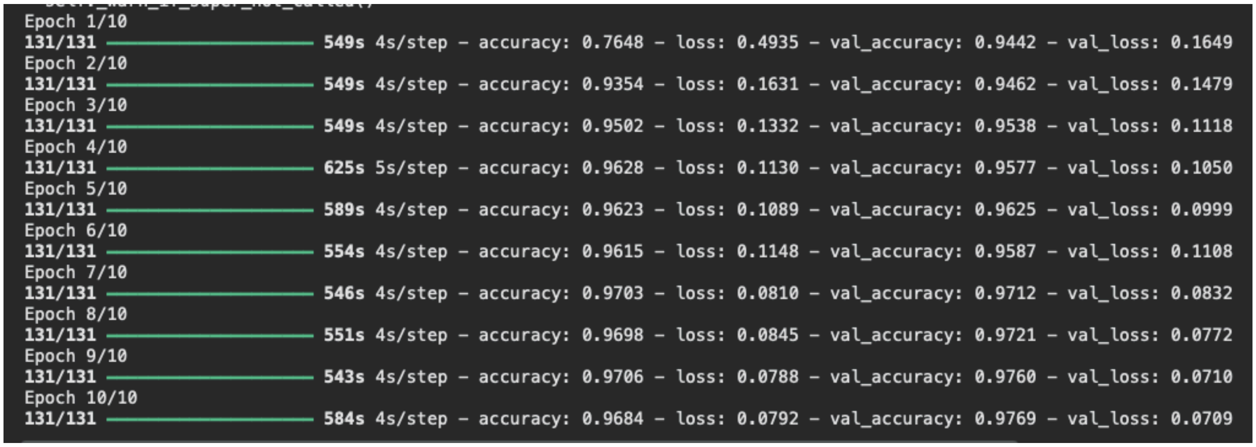
Receiver Operating Characteristic (ROC) curves and **Area Under the Curve (AUC)**: Used to evaluate model performance across varying classification thresholds, providing insight into the trade-off between sensitivity and specificity.

These metrics were computed on the held-out test set for all models to ensure consistent and unbiased performance comparison.

4.6 Results and Evaluations

This section presents the experimental results obtained from all three models. Performance is evaluated on the held-out test set using the metrics described in the previous section. Results are presented using confusion matrices, ROC curves, and quantitative summaries to allow for direct comparison across models.

4.6.1 CNN Results



```
Epoch 1/10
131/131 ━━━━━━━━━━━━━━━━━━━━ 549s 4s/step - accuracy: 0.7648 - loss: 0.4935 - val_accuracy: 0.9442 - val_loss: 0.1649
Epoch 2/10
131/131 ━━━━━━━━━━━━━━━━━━━━ 549s 4s/step - accuracy: 0.9354 - loss: 0.1631 - val_accuracy: 0.9462 - val_loss: 0.1479
Epoch 3/10
131/131 ━━━━━━━━━━━━━━━━━━━━ 549s 4s/step - accuracy: 0.9502 - loss: 0.1332 - val_accuracy: 0.9538 - val_loss: 0.1118
Epoch 4/10
131/131 ━━━━━━━━━━━━━━━━━━━━ 625s 5s/step - accuracy: 0.9628 - loss: 0.1130 - val_accuracy: 0.9577 - val_loss: 0.1050
Epoch 5/10
131/131 ━━━━━━━━━━━━━━━━━━━━ 589s 4s/step - accuracy: 0.9623 - loss: 0.1089 - val_accuracy: 0.9625 - val_loss: 0.0999
Epoch 6/10
131/131 ━━━━━━━━━━━━━━━━━━━━ 554s 4s/step - accuracy: 0.9615 - loss: 0.1148 - val_accuracy: 0.9587 - val_loss: 0.1108
Epoch 7/10
131/131 ━━━━━━━━━━━━━━━━━━━━ 546s 4s/step - accuracy: 0.9703 - loss: 0.0810 - val_accuracy: 0.9712 - val_loss: 0.0832
Epoch 8/10
131/131 ━━━━━━━━━━━━━━━━━━━━ 551s 4s/step - accuracy: 0.9698 - loss: 0.0845 - val_accuracy: 0.9721 - val_loss: 0.0772
Epoch 9/10
131/131 ━━━━━━━━━━━━━━━━━━━━ 543s 4s/step - accuracy: 0.9706 - loss: 0.0788 - val_accuracy: 0.9760 - val_loss: 0.0710
Epoch 10/10
131/131 ━━━━━━━━━━━━━━━━━━━━ 584s 4s/step - accuracy: 0.9684 - loss: 0.0792 - val_accuracy: 0.9769 - val_loss: 0.0709
```

Figure 3: CNN training log over 10 epochs from google colab showing training and validation accuracy and loss statistics.

Shows consistent improvement across epochs, indicating effective learning. Validation performance closely follows training performance, suggesting limited overfitting. Minor divergence observed in later epochs motivated a reduction in the number of training epochs to improve generalization.

Table 2: Confusion matrix for the CNN model on the test set.

Actual/Predicted	Normal	Pneumonia
Normal	97	137
Pneumonia	3	387

Table 3: CNN classification metrics on the test set.

Class	Precision	Recall	F1-Score
Normal	0.97	0.41	0.58
Pneumonia	0.74	0.99	0.85
Overall Accuracy			0.78

The CNN model shows strong sensitivity toward pneumonia cases, achieving a recall of 0.99 for the positive class. This indicates that nearly all pneumonia cases were correctly identified. However, the recall for the normal class is considerably lower (0.41), meaning that many normal images were misclassified as pneumonia.

This behavior reflects a trade-off favoring sensitivity over specificity, which is common in medical imaging tasks and is further influenced by the class imbalance in the dataset. While overall accuracy is 0.78, this result reinforces the importance of using precision, recall, and F1-score rather than accuracy alone when evaluating performance on imbalanced datasets.

In addition to standard classification metrics, confusion-matrix–based measures were examined. The CNN achieved a high **True Positive Rate (TPR)(0.991)**, indicating strong sensitivity to pneumonia cases, while the **True Negative Rate (TNR) (0.415)** was lower, reflecting increased false positives for normal cases. **The Negative Predictive Value (NPV) (0.970)** remained high, suggesting that predictions labeled as normal were generally reliable.

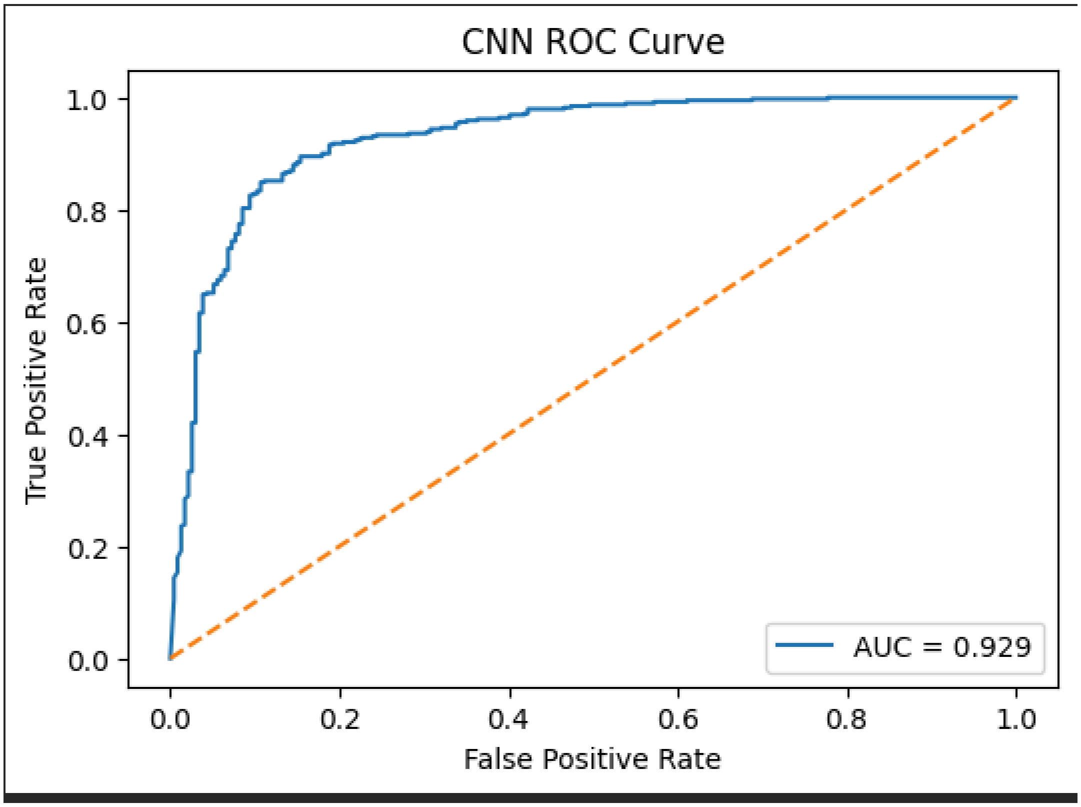
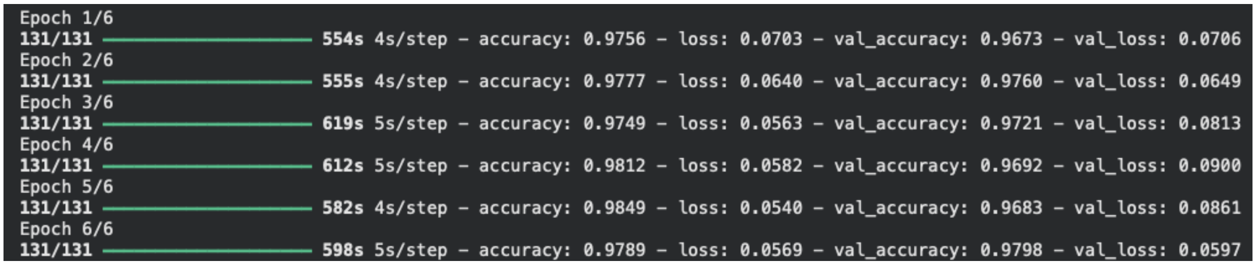


Figure 4: ROC curve for the CNN model on the test dataset.

The ROC curve further demonstrates the CNN's discriminative ability, achieving a high AUC value, which indicates strong separation between pneumonia and normal cases across decision thresholds.

Overfitting: Training and validation performance were monitored across multiple epochs. While training accuracy continued to improve with additional epochs, validation accuracy plateaued after approximately six epochs. This behavior suggests the onset of mild overfitting beyond this point. As a result, a reduced-epoch model was also evaluated to demonstrate the trade-off between model complexity and generalization.



```
Epoch 1/6
131/131 ━━━━━━━━━━━━━━━━━━━━ 554s 4s/step - accuracy: 0.9756 - loss: 0.0703 - val_accuracy: 0.9673 - val_loss: 0.0706
Epoch 2/6
131/131 ━━━━━━━━━━━━━━━━━━━━ 555s 4s/step - accuracy: 0.9777 - loss: 0.0640 - val_accuracy: 0.9760 - val_loss: 0.0649
Epoch 3/6
131/131 ━━━━━━━━━━━━━━━━━━━━ 619s 5s/step - accuracy: 0.9749 - loss: 0.0563 - val_accuracy: 0.9721 - val_loss: 0.0813
Epoch 4/6
131/131 ━━━━━━━━━━━━━━━━━━━━ 612s 5s/step - accuracy: 0.9812 - loss: 0.0582 - val_accuracy: 0.9692 - val_loss: 0.0900
Epoch 5/6
131/131 ━━━━━━━━━━━━━━━━━━━━ 582s 4s/step - accuracy: 0.9849 - loss: 0.0540 - val_accuracy: 0.9683 - val_loss: 0.0861
Epoch 6/6
131/131 ━━━━━━━━━━━━━━━━━━━━ 598s 5s/step - accuracy: 0.9789 - loss: 0.0569 - val_accuracy: 0.9798 - val_loss: 0.0597
```

Figure 5: Training and Validation loss across 6 epochs in CNN showing training, accuracy and validation loss.

The model generalizes well to unseen data and that overfitting is limited when training is stopped at six epochs. Compared to longer training runs, the six-epoch configuration provides a better balance between model complexity and generalization.

4.6.2 Logistic Regression Results

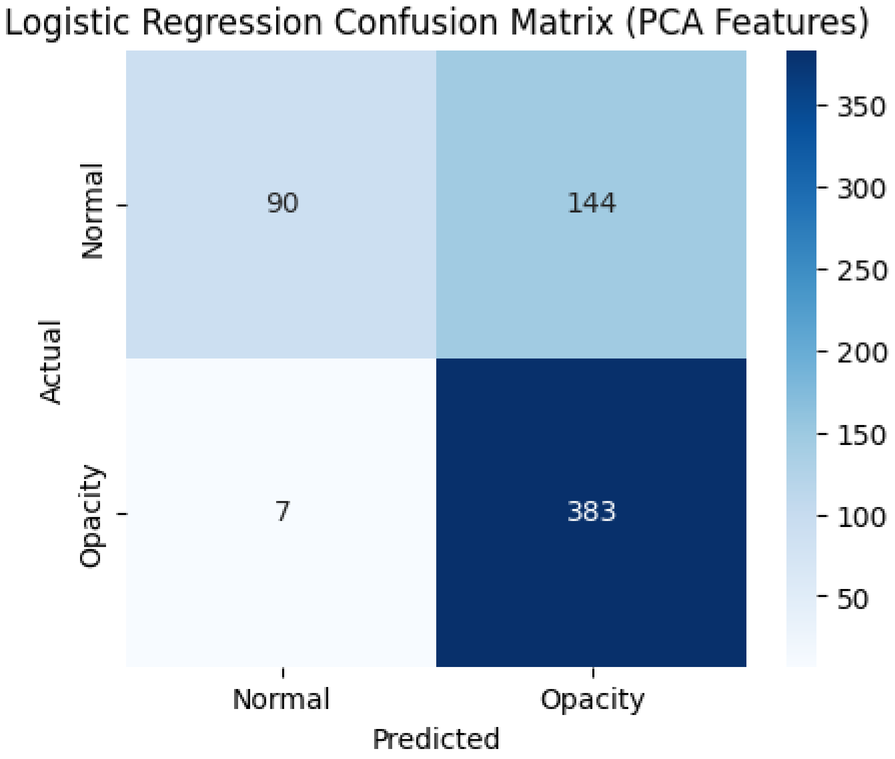


Figure 6: Confusion matrix for the PCA + Logistic Regression model on the test set.

The confusion matrix shows that the Logistic Regression model correctly identified the majority of pneumonia cases, achieving high sensitivity for the positive class. However, a substantial number of normal cases were misclassified as pneumonia, indicating a bias toward the majority class. This behavior is consistent with the imbalanced nature of the dataset.

Name of the Metric	Value
Accuracy	0.76
Precision	0.73
Recall/TPR	0.98
f1-Score	0.84
TNR (class 0)	0.38
NPV	0.93

Table 4: Classification metrics for the PCA + Logistic Regression model on the test dataset.

The Logistic Regression model achieved an overall accuracy of approximately 0.76. Recall for pneumonia cases was high (0.98), demonstrating strong sensitivity, while recall for normal cases was significantly lower (0.38). Precision and F1-score values reflect the trade-off between correctly identifying pneumonia cases and an increased false positive rate for normal cases.

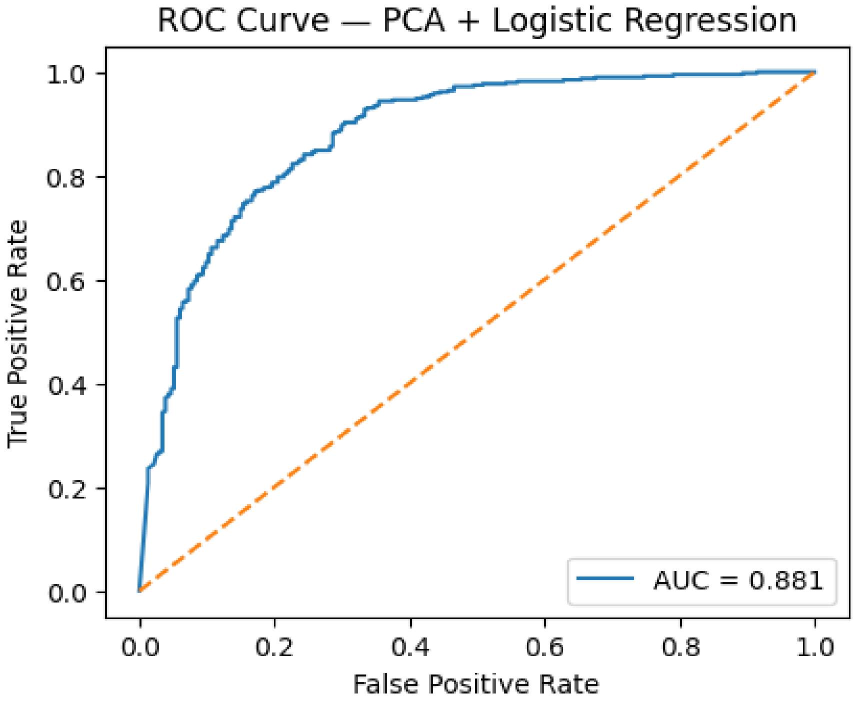


Figure 7: ROC curve for the PCA + Logistic Regression model on the test dataset.

The ROC curve demonstrates strong discriminative performance, with an AUC of approximately 0.88. This indicates that the model is able to separate pneumonia and normal cases effectively across different classification thresholds despite its linear structure.

Overfitting: Logistic Regression does not involve iterative epoch-based training and therefore does not exhibit traditional overfitting behavior. Overfitting was mitigated through PCA-based dimensionality reduction and L2 regularization. Validation and test performance remained consistent, suggesting reasonable generalization. However, the linear nature of the model limits its ability to capture complex spatial patterns present in chest X-ray images.

4.6.3 Random Forest Results

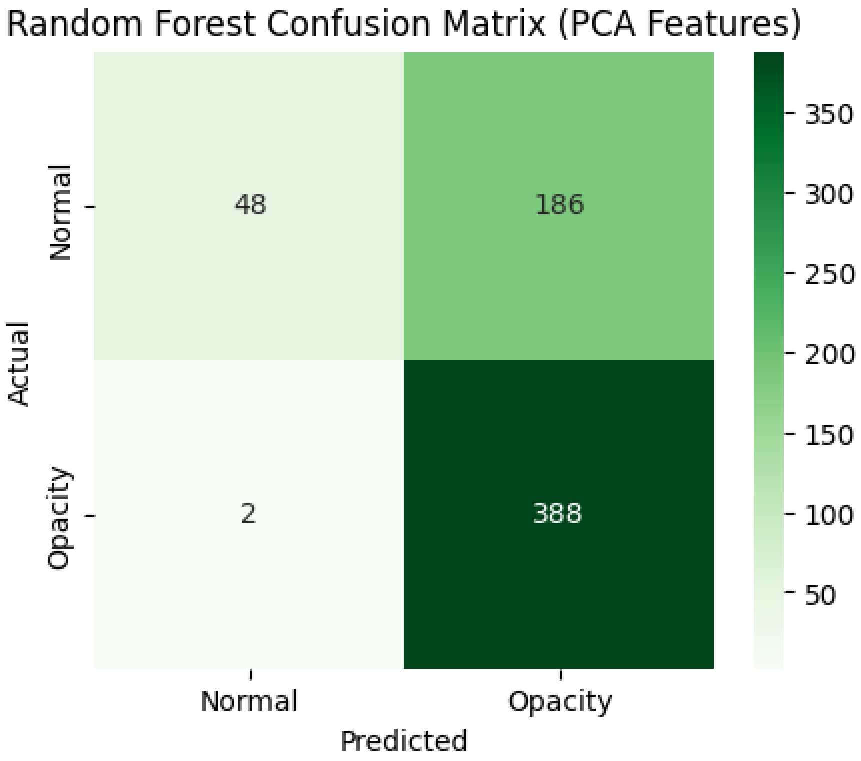


Figure 8: Confusion matrix for the Random Forest model on test dataset

The confusion matrix shows that the Random Forest model correctly identifies the majority of pneumonia cases while misclassifying a significant number of normal cases as pneumonia. This indicates a strong bias toward the positive class, similar to the behavior observed in the CNN and Logistic Regression models.

Name of the Metric	Value
Accuracy	0.70
Precision	0.68
Recall/TPR	0.99
f1-Score	0.80
TNR (class 0)	0.21
NPV	0.96

Table 5: Classification metrics for the Random Forest model on the test dataset

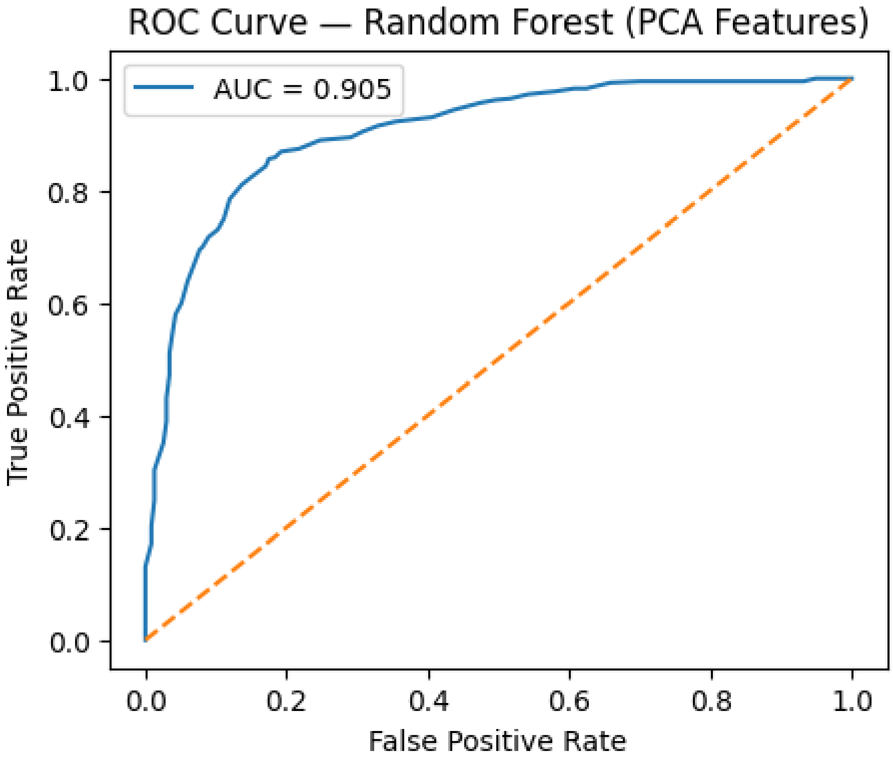


figure 9: ROC curve for the Random Forest Model on the test dataset

The ROC curve demonstrates strong discriminative ability, with an AUC of approximately **0.90**, indicating effective class separation across decision thresholds despite reduced overall accuracy.

Overfitting: Random Forest models reduce overfitting by averaging predictions across multiple decision trees. However, due to the imbalance in the dataset and the model's emphasis on sensitivity, the classifier exhibits reduced specificity for normal cases. While overfitting was partially mitigated through ensemble learning, the model's performance suggests a trade-off between sensitivity and specificity similar to a pattern that other models have shown too.

4.7 Model Comparison

Model	**Accuracy**	**Precision (Pneumonia)**	**Recall (Pneumonia)**	**F1-Score (Pneumonia)**	**ROC AUC**
CNN	0.78	0.74	0.99	0.85	0.92
Logistic Regression (PCA)	0.76	0.73	0.98	0.84	0.88
Random Forest	0.70	0.68	0.99	0.80	0.90

Table 6: Performance comparison across evaluated models

Table 6 compares the performance of the three evaluated models across multiple metrics. **The CNN** achieved the highest overall accuracy and F1-score while maintaining strong sensitivity to pneumonia cases, demonstrating its ability to capture complex spatial patterns in chest X-ray images.

Logistic Regression with PCA performed competitively despite its linear structure, achieving a high ROC AUC and strong recall for pneumonia cases, but with reduced specificity.

The Random Forest model exhibited the highest sensitivity among the classical models but suffered from lower overall accuracy and specificity, indicating a stronger bias toward the majority class.

4.7.1 Sensitivity and Specificity Trade-Offs

Across all models, a consistent trade-off between sensitivity and specificity was observed. All three models favored the detection of pneumonia cases, resulting in high recall values for the positive class but lower recall for normal cases. This behavior is influenced by the class imbalance in the dataset and reflects a common design choice in medical imaging tasks, where minimizing false negatives is often prioritized over minimizing false positives.

4.7.2 Model Complexity and Computational Considerations

CNN required the highest computational cost due to iterative training and large parameter space but provided the strongest overall performance. Logistic Regression with PCA was computationally efficient and interpretable, making it a suitable baseline model. Random Forest provided nonlinear decision boundaries but exhibited reduced specificity, suggesting limited effectiveness when applied directly to high-dimensional image features.

4.8 Discussion

4.8.1 Impact of Class Imbalance

A key characteristic of the dataset used in this project is the imbalance between pneumonia and normal chest X-ray images. Pneumonia cases occur more frequently than normal cases, which influenced the behavior of all evaluated models. Rather than correcting this imbalance through resampling or class weighting, the original class distribution was preserved to reflect realistic clinical data. As a result, all models exhibited a bias toward predicting pneumonia, leading to high sensitivity for the positive class but reduced specificity for normal cases.

This outcome highlights the importance of evaluating performance using metrics beyond accuracy. Precision, recall, F1-score, and class-specific measures provided a more informative assessment of model behavior under imbalanced conditions.

4.8.2 Overfitting and Generalization

Overfitting behavior differed across models due to differences in model complexity and training mechanisms. The CNN exhibited mild overfitting when trained for an extended number of epochs, as evidenced by divergence between training and validation performance. Reducing the number of training epochs improved generalization while maintaining strong classification performance.

In contrast, Logistic Regression did not exhibit traditional overfitting behavior due to its linear structure and lack of iterative training. Overfitting was mitigated through PCA-based dimensionality reduction and L2 regularization. Random Forest reduced overfitting through ensemble averaging but still demonstrated reduced specificity, likely due to the dataset imbalance and feature representation.

4.8.3 Model Limitations

Model limitations should be acknowledged to better all future work. All models were trained on resized and simplified image representations, which may discard fine grained anatomical details. Additionally, limited hyperparameter tuning was performed, as the focus of this project was on model comparison and interpretation rather than optimization. The dataset represents a specific clinical context, and model performance may not generalize to different imaging protocols or patient populations.

4.8.4 Future Work

Future work could explore advanced preprocessing techniques, such as data augmentation or class-weighted loss functions, to address class imbalance more effectively. More extensive hyperparameter tuning and the inclusion of additional deep learning architectures could further improve performance. Incorporating explainability techniques, such as saliency maps, could also enhance model interpretability in clinical settings.

4.9 Conclusion

This project explored the application of multiple machine learning approaches to the task of classifying chest X-ray images as normal or pneumonia. Using an imbalanced medical imaging dataset, three different models were evaluated: a Convolutional Neural Network (CNN), a classical Logistic Regression model with PCA, and an ensemble-based Random Forest classifier. The objective was not to achieve optimal predictive performance, but rather to understand model behavior, evaluate performance under class imbalance, and interpret results using a variety of evaluation metrics.

The CNN demonstrated the strongest overall performance, achieving high accuracy, strong F1-scores, and excellent sensitivity for pneumonia cases. Its ability to learn spatial features directly from image data allowed it to outperform the classical models, though careful monitoring of overfitting was required. Logistic Regression with PCA provided a competitive baseline, achieving strong ROC–AUC performance despite its linear structure, highlighting the effectiveness of dimensionality reduction in high dimensional image data. The Random Forest model achieved high sensitivity but suffered from reduced specificity, emphasizing the trade-offs associated with ensemble methods when applied to imbalanced datasets.

Across all models, a consistent pattern emerged: high sensitivity to pneumonia cases paired with lower specificity for normal cases. This reflects both the class imbalance present in the dataset and the clinical priority of minimizing false negatives. The use of multiple evaluation metrics including precision, recall, F1-score, confusion matrices, and ROC–AUC proved essential for accurately interpreting model performance beyond overall accuracy.

Overall, this project demonstrates that different machine learning models exhibit distinct strengths and limitations when applied to medical imaging tasks. While deep learning approaches offer superior performance, classical models remain valuable for baseline comparison and interpretability. The results underscore the importance of careful metric selection, transparent evaluation, and thoughtful discussion of limitations when applying machine learning in healthcare contexts.